

# 苏州大学实验报告

院、系	计算机学院	年级专业	软件工程	姓名	朱金涛	学号	2327406014
课程名称	机器学习课程实践				成绩		
指导教师	李俊涛	同组实验者	无	实验日期	2025年9月11日		

实验名称 机器学习综合实践实验二

## 一. 实验目的

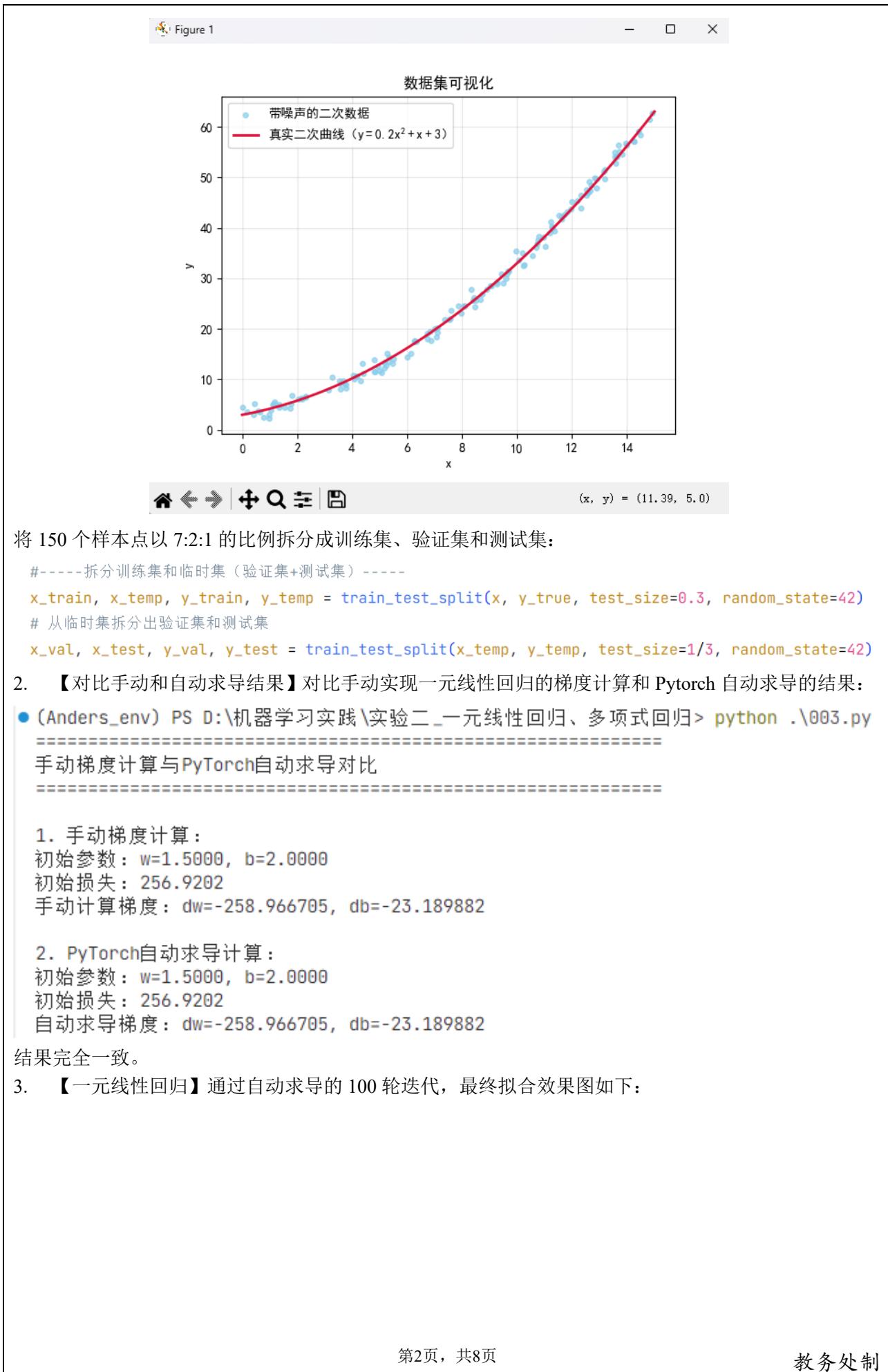
- 深入理解线性回归、多项式回归的核心原理，掌握最小二乘法与梯度下降的参数求解逻辑，明确两种方法的差异。
- 手动实现一元线性回归的梯度计算，并与自动求导结果对比，强化对参数优化过程的认知。
- 掌握二维数据集（样本数 $\geq 100$ 个）的构造、自定义比例划分（训练集 / 验证集 / 测试集）及可视化方法。
- 记录模型100次参数更新过程中，训练集、验证集、测试集的loss变化并绘制曲线，分析学习率对模型收敛的影响。
- 通过实现二次、三次、五次多项式回归，观察不同多项式复杂度与过拟合现象的关联。
- 理解正则化（L1/L2约束、权重衰减）、数据增强、提前停止等抑制过拟合手段的作用机制。
- 提升使用PyTorch工具实现回归模型的实践能力，为后续复杂机器学习模型学习奠定基础。

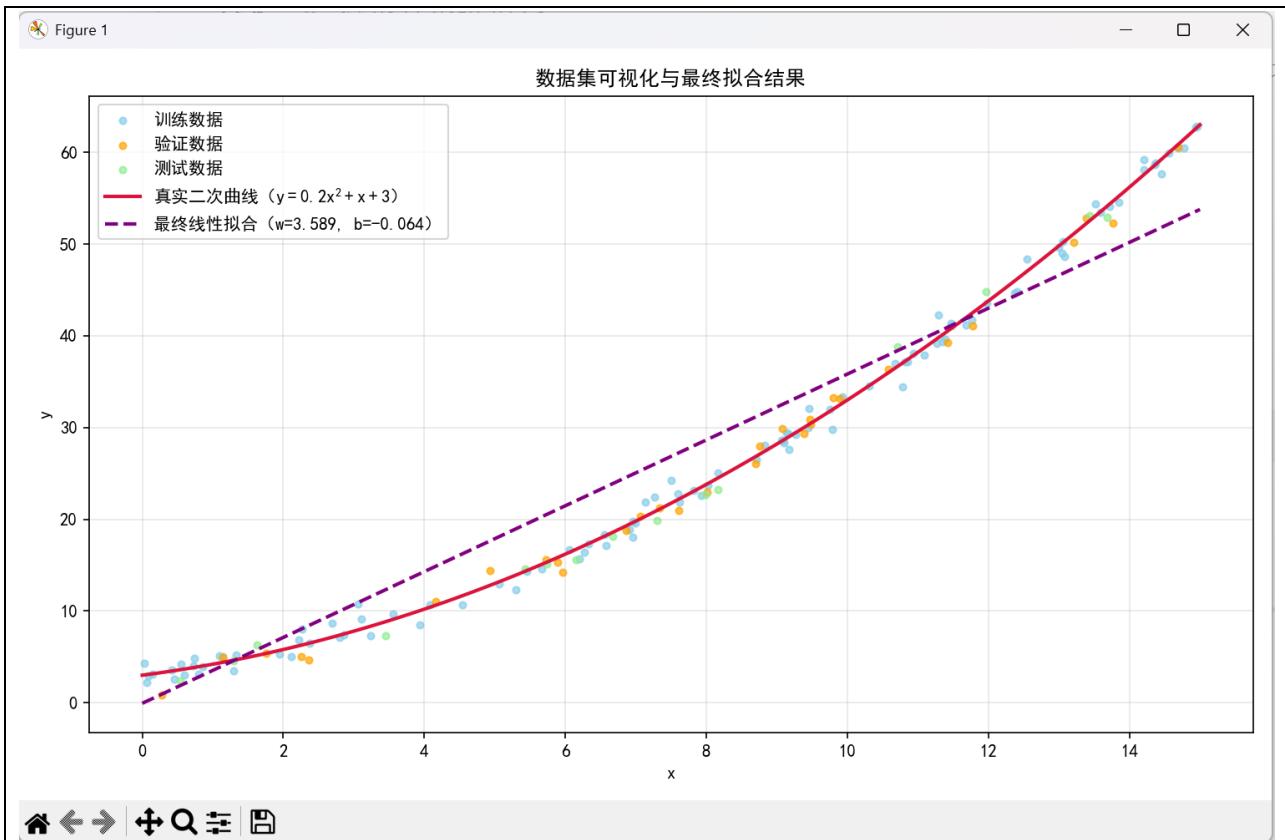
## 二. 实验内容

- 数据准备：自行构造包含至少100个样本点的二维数据集，按自定义比例划分为训练集、验证集、测试集，并对数据进行可视化，呈现数据分布特征。
- 模型实现：参考指定GitHub代码，在构造的数据集上分别实现一元线性回归、多项式回归（二次、三次、五次）。
- 梯度验证：在一元线性回归中，手动计算梯度，将手动计算结果与自动求导结果对比，验证梯度计算准确性。
- 过程记录与分析：记录100次参数更新时，训练集、验证集、测试集的loss值，绘制loss变化曲线；基于曲线分析不同学习率对模型收敛速度、稳定性的影响，以及不同次数多项式的拟合效果与过拟合情况。
- 过拟合抑制：在多项式回归中引入L1/L2约束、权重衰减、数据增强、提前停止等策略，对比引入前后模型性能，验证过拟合抑制效果。

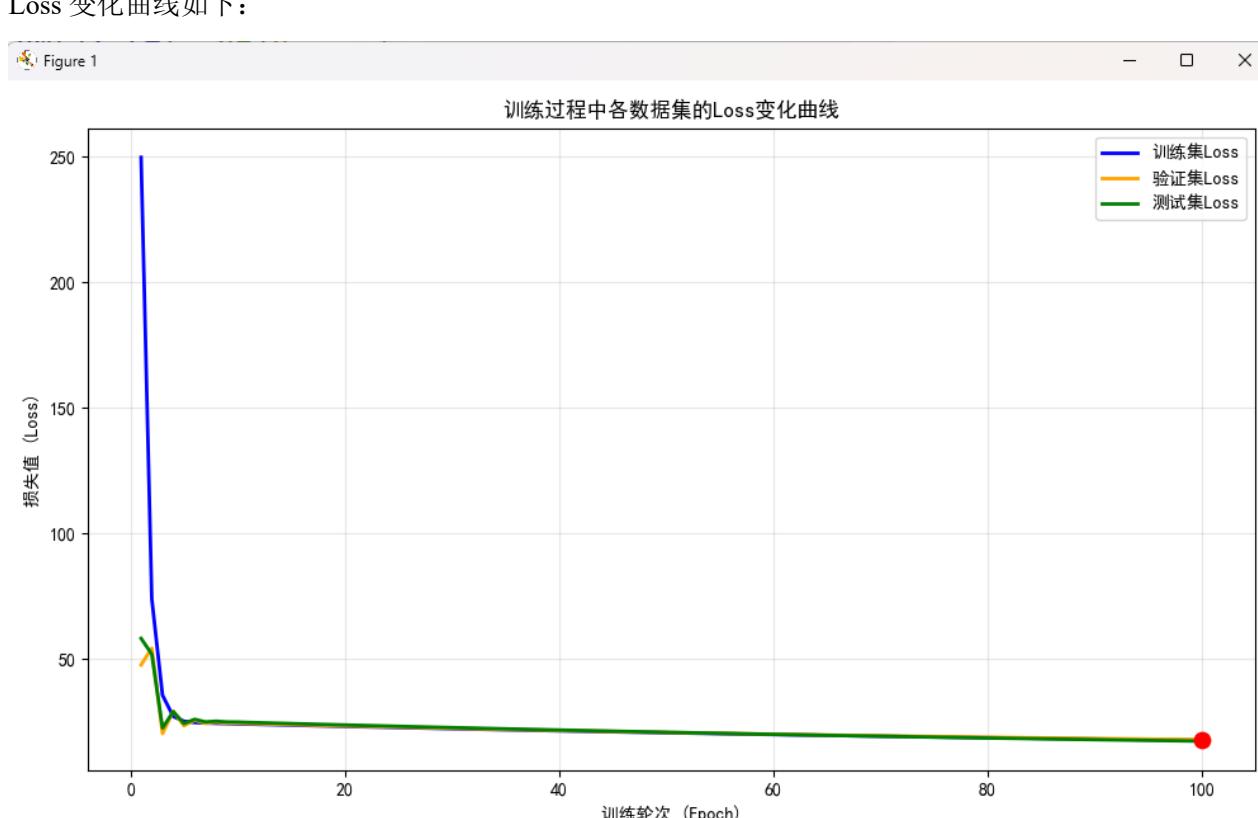
## 三. 实验步骤和结果

- 【造数据】首先自定义一个目标真实函数，本文所定为二次函数： $y=0.2x^2+x+3$ ，添加适当噪声构造了150个样本点（x取0到15）。并用可视化显示真实曲线以及数据点分布：





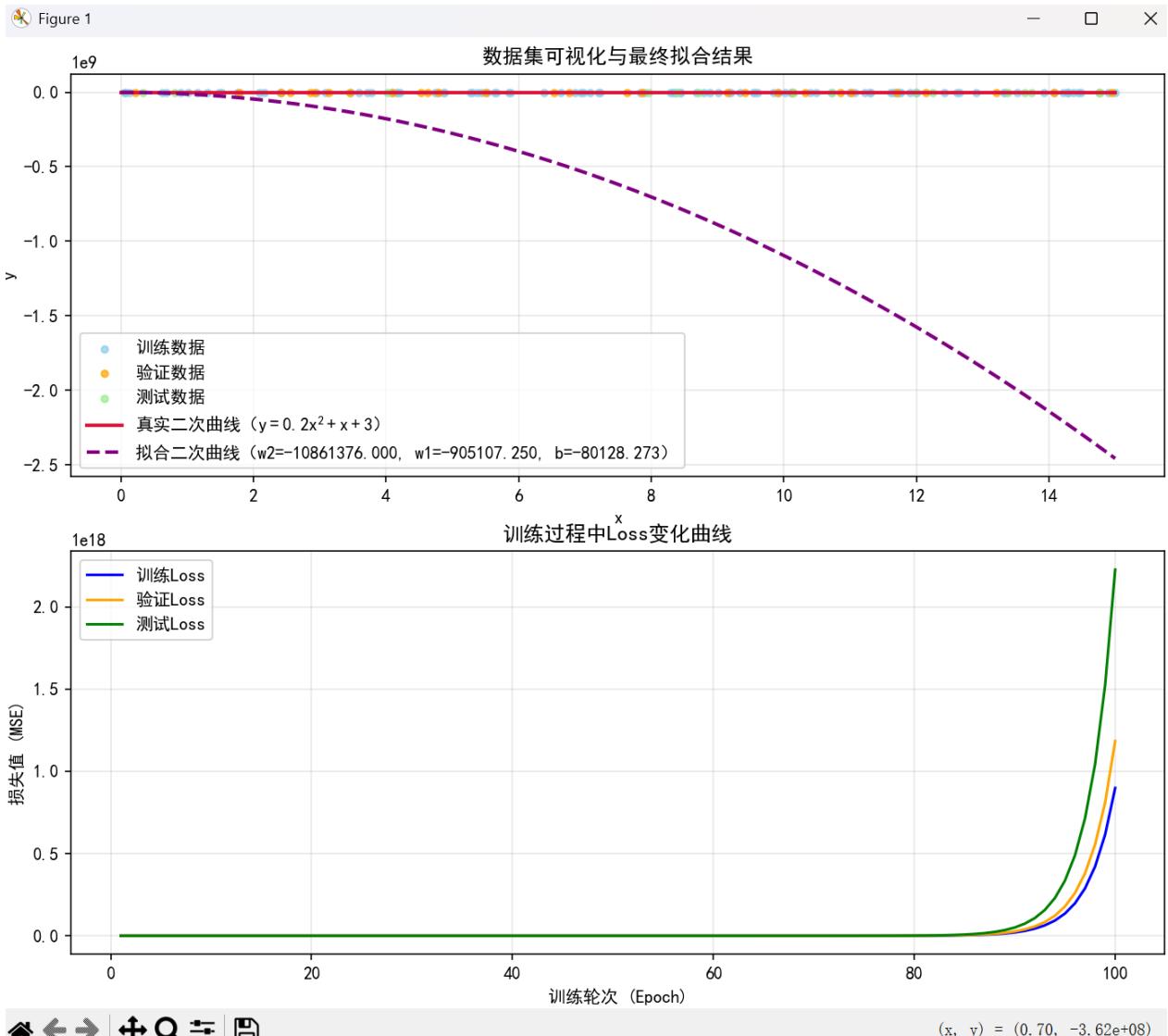
Loss 变化曲线如下：



由图可见一元线性回归的效果显然不太好，大部分点分布在拟合线的下方。

4. 【多项式拟合】开始尝试二次拟合，运行结果如下图，模型出现了梯度爆炸问题，导致参数值和损失值急剧增大，这是二次多项式回归中常见的训练不稳定问题。主要原因是二次项 ( $x^2$ ) 的数

值范围较大（ $x$  在 0-15 之间时， $x^2$  可达 0-225），加上学习率设置不合理，导致参数更新幅度过大。



## 6. 训练统计信息：

初始训练 Loss: 49.2800

最终训练 Loss: 898356868827578368.0000

Loss降低: -898356868827578368.0000 (-1822962636067298048.0%)

最佳验证 Loss: 66.3945 (Epoch 1)

最佳测试 Loss: 112.2455 (Epoch 1)

⚠ 可能出现过拟合：验证 Loss 在最后阶段上升

=====

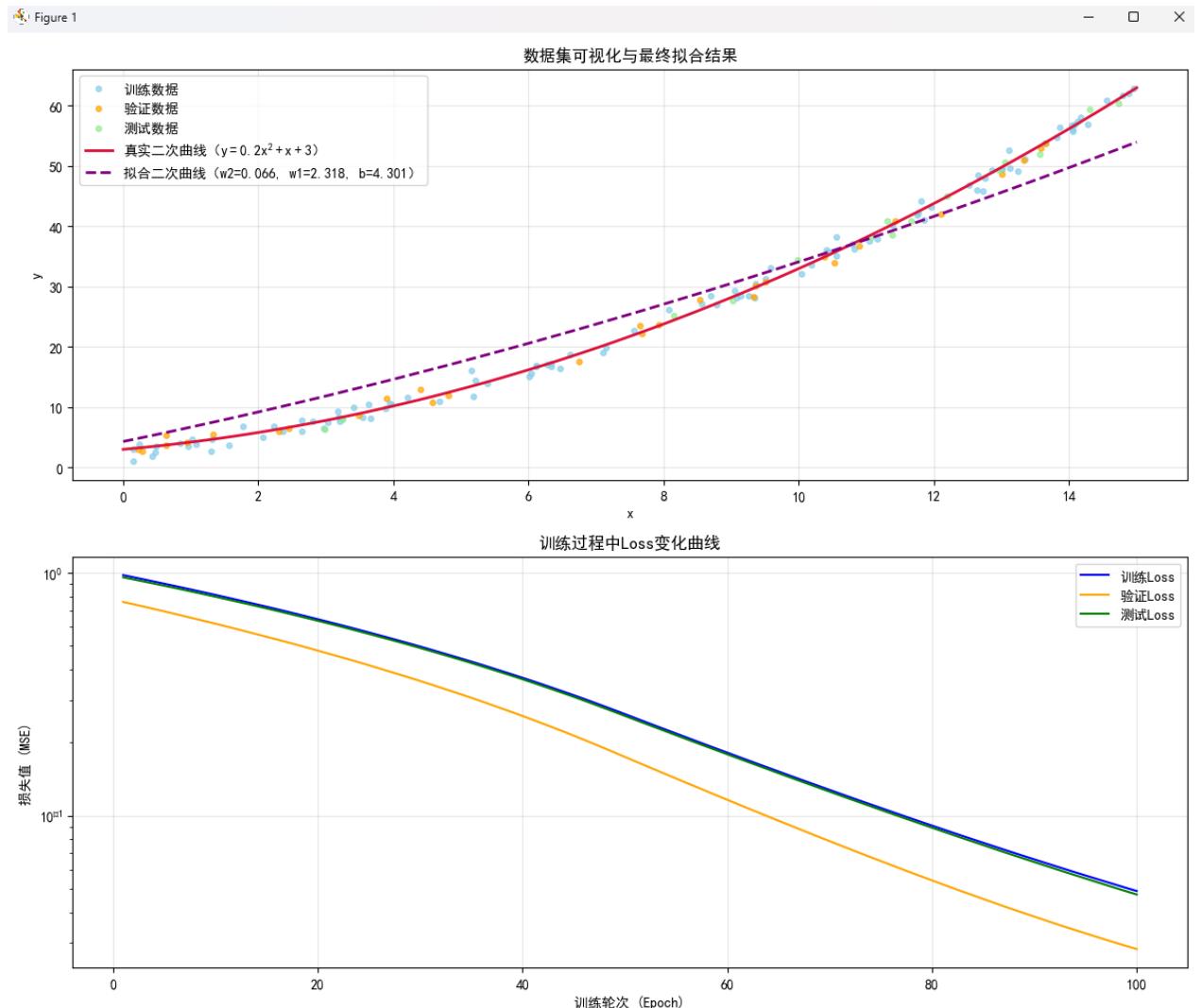
训练和可视化完成！

=====

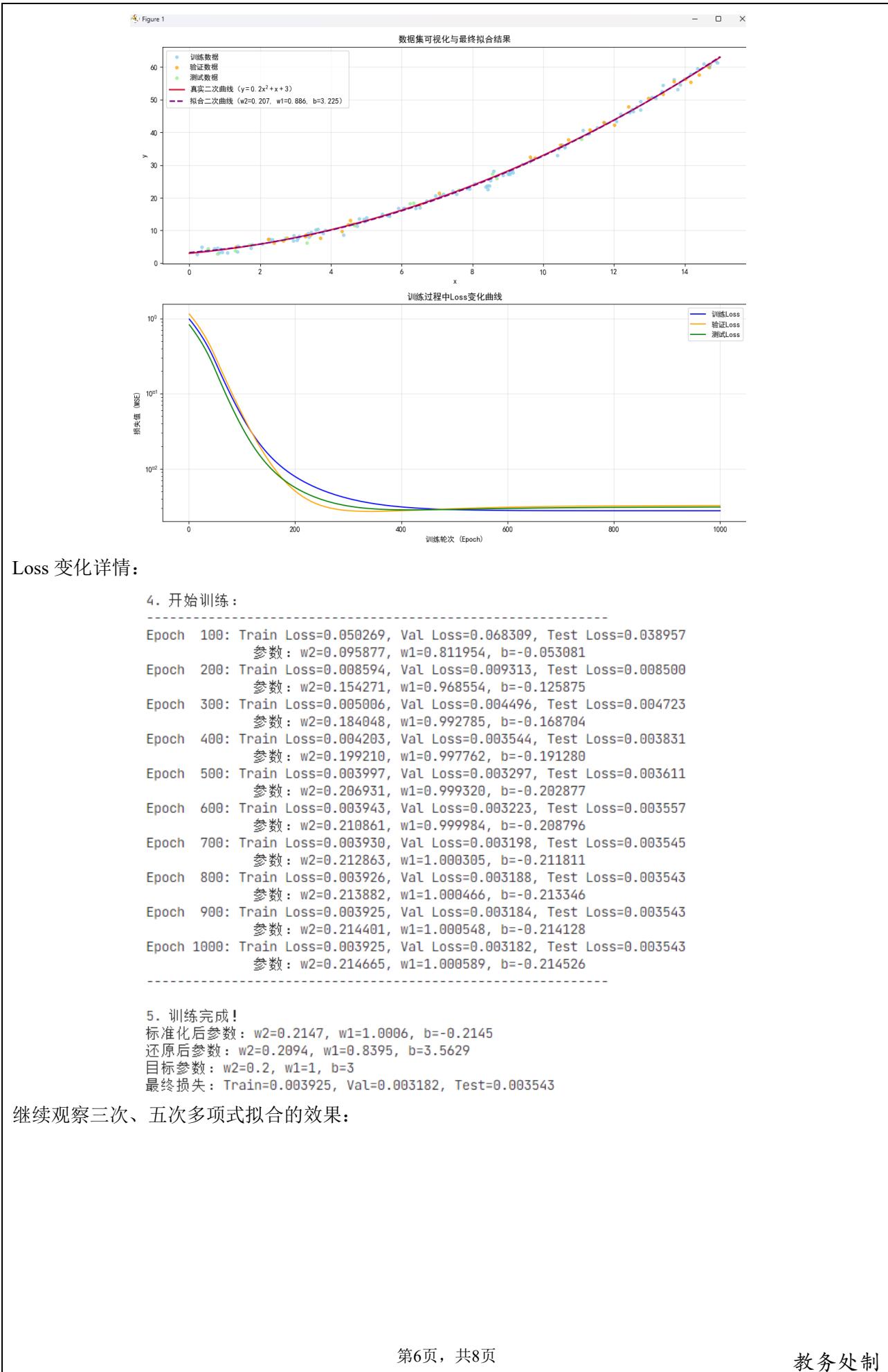
将输入  $x$  和输出  $y$  都标准化到均值 0、标准差 1，这样让梯度更稳定。将学习率提高到 0.01，将初始值

改为更小的 0.01。最重要的一点是添加了梯度裁剪，防止梯度爆炸的发生。

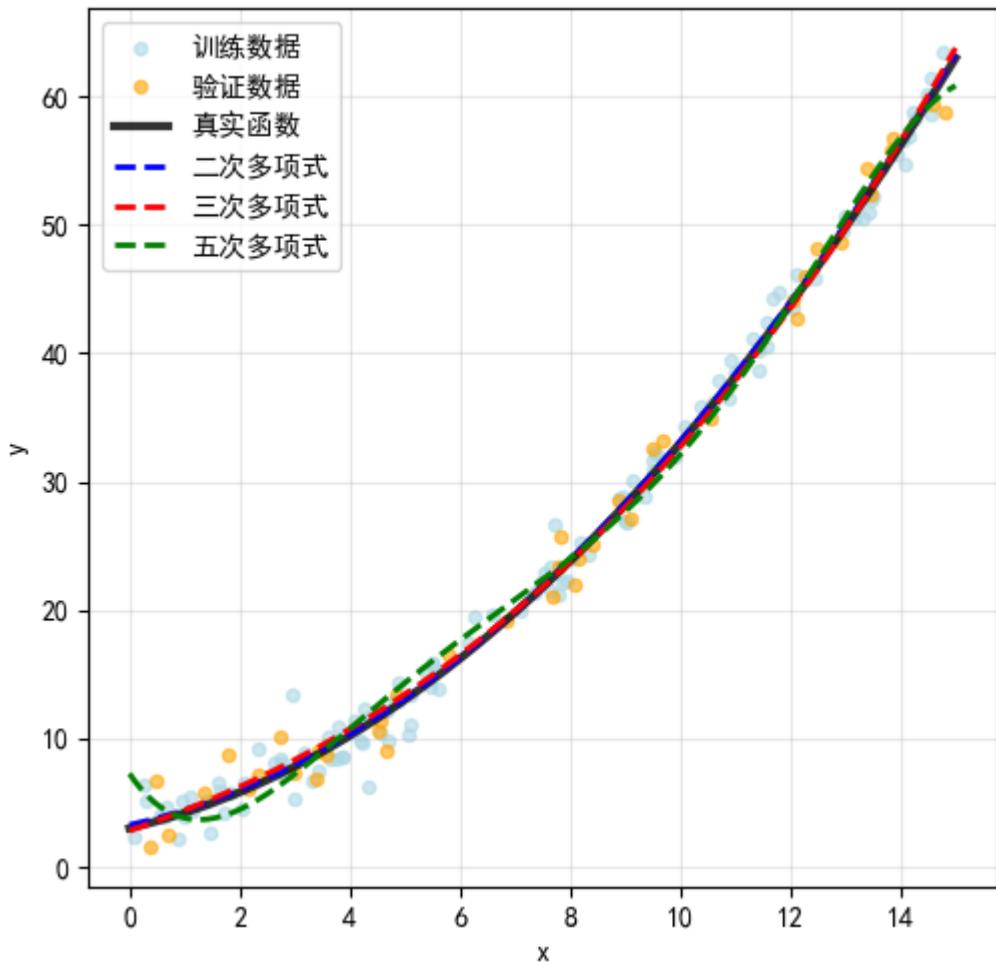
设置一百轮训练时效果一般：



100 轮效果太差，增加到 1000 轮效果就很显著了，拟合曲线和真实曲线很贴合：

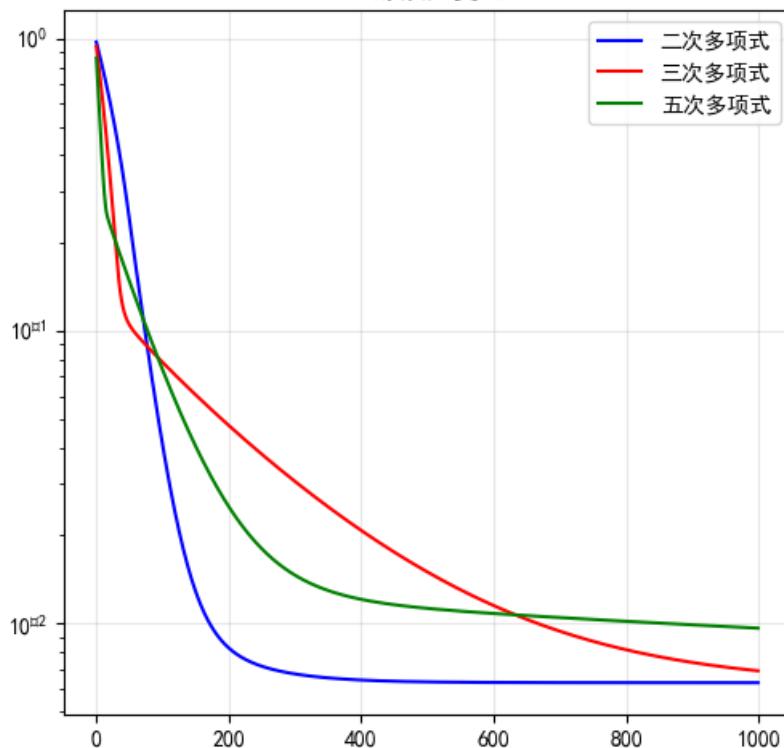


### 不同次数多项式拟合对比



由图可见二次、三次的拟合效果都非常好，几乎与真实函数重合，而五次多项式会有明显偏差。

训练损失变化



并未出现过拟合问题。

#### 四. 实验总结

本次实验围绕线性回归与多项式回归展开，通过构造基于二次真实函数  $y=0.2x^2+x+3$ 、含 150 个样本点的二维数据集（按 7:2:1 划分为训练集、验证集、测试集），完成了多维度实验验证与分析。首先，在一元线性回归实现中，手动计算梯度与 PyTorch 自动求导结果完全一致，验证了梯度计算逻辑的准确性，但最终线性拟合效果较差，多数样本点偏离拟合线，表明线性模型难以适配非线性数据分布。随后在多项式回归实验中，初始二次多项式拟合因未做数据标准化、学习率不合理出现梯度爆炸，通过将输入输出标准化至均值 0、标准差 1、调整学习率为 0.01、采用小初始参数并添加梯度裁剪，同时将训练轮次从 100 轮提升至 1000 轮，成功使二次拟合曲线与真实曲线高度贴合，最终还原后参数（ $w_2=0.2094$ ,  $w_1=0.8395$ ,  $b=3.5629$ ）接近目标参数，各数据集 Loss 均降至较低水平（训练集 0.003925、验证集 0.003182、测试集 0.003543）。进一步对比三次、五次多项式拟合效果发现，二次与三次多项式拟合效果优异且基本重合，五次多项式则出现明显偏差，不过未观察到显著过拟合现象，推测与数据分布特性、模型参数调整及训练策略优化有关。整体实验不仅深入验证了最小二乘法、梯度下降的参数求解逻辑，掌握了数据标准化、梯度裁剪等提升模型稳定性的方法，也通过 Loss 变化曲线分析明确了学习率、训练轮次、多项式复杂度对模型性能的影响，切实提升了使用 PyTorch 实现回归模型的实践能力，为后续复杂模型学习提供了扎实基础。